Original Manuscript ID: futureinternet-3515366

Original Article Title: "Multimodal Fall Detection Using Spatial-Temporal Attention and Bi-LSTM
Based Feature Fusion"

To Future Internet Editor

Re: Response to reviewers

Dear Editor,

Thank you for allowing a resubmission of our manuscript and allowing us to address the reviewers' comments. We are uploading (a) our point-by-point response to the comments (below) (response to reviewers, under

"Author's Response Files"), (b) an updated manuscript with yellow highlighting indicating changes (as

"Highlighted PDF"), and (c) a clean, updated manuscript without highlights ("Main Manuscript").

Best regards,

<**J. Shin**> et al.

Reviewer #1

This paper investigates multimodal fall detection using spatial-temporal attention and Bi-LSTM-based feature fusion. The overall architecture consists of two streams, with the first stream being used to capture spatial-temporal relationships using Graph-based Spatial-Temporal Convolutional and Attention Neural Networks, while the second stream applies Bi-LSTM on accelerometer data. The outputs of both streams are concatenated and fed into a fully connected layer for classification. The system is evaluated on the Fall Up dataset and achieves 99.09% classification accuracy. Although the method achieves good results, several concerns need to be addressed.

1) The system has limited novelty as it combines the already proposed model GSTCN with BILSTM, which others have proposed. The novelty lies in only using the multimodal data, however, the intuition behind using accelerometer data is missing. Please explain in paper why you chose only accelerometer data instead of other in the UP Fall dataset.

   Author response: Thank you very much for raising your concern. We acknowledge that both the GSTCAN and Bi-LSTM models have been previously proposed in the literature. However, the novelty of our work lies not only in combining these models but also in integrating them with multimodal data to improve fall detection accuracy. The combination of skeleton and accelerometer data allows us to capture both spatial-temporal relationships and long-range dependencies, providing a more comprehensive understanding of human motion.

   Regarding the use of accelerometer data from the UP Fall dataset, we selected it because accelerometers provide rich, reliable, and continuous motion information, especially for fall detection tasks. They are sensitive to abrupt movements and can capture dynamic changes in orientation and velocity, which are crucial for detecting falls. Moreover, accelerometers are commonly used in wearable devices, making them practical for real-world healthcare applications.

2) As GSTCAN is proposed by [46], the authors may not explain this model in so much detail as it is not proposed by this paper. Or else give new analysis on this model that will be interesting to readers and also write if authors do anything different from [46].

   Author response: Thank you very much for raising your concern. We appreciate your suggestion and understand the need to avoid excessive detail on the GSTCAN model, especially since it was originally proposed by [46]. Regarding the hyperparameter tuning of the GSTCAN, we will provide a clearer explanation of how this process was carried out in our work and highlight any modifications or optimisations that differentiate our approach from [46]. While the foundational structure of GSTCAN remains the same, we focused on hyperparameter tuning in our work. Specifically, we experimented with 3, 4, 6, and 9 GSTCAN modules and achieved the best accuracy with 9 GSTCAN modules. We also observed that the results with 4, 5, and 6 GSTCAN modules were quite similar. Specifically, Table 3, the 5 GSTCAN module achieved 99.24, and the 6 GSTCAN module produced 99.16. We suggest that in paper [46], they could have used 5 GSTCAN modules instead of 6 to reduce computational complexity.

Our novelty lies in the use of varying numbers of GSTCAN modules, integration with accelerometer data, the fusion of multimodal features, and the application of a graph convolutional network with a Channel Attention (CA) mechanism, which was not explored in the original GSTCAN paper [46]. We included ablation studies in Tables 2 and 3 for the Fall-up and UR-Fall datasets, showing performance accuracy with multimodal datasets across various Sensor and Skeleton datasets configurations. Specifically, the number of GSTCAN modules clearly distinguishes between [46] and our work.

Table 2. Ablation study of the proposed model for UP-Fall multi-modal dataset.

Ablation	Stream-1 Skeleton		Stream-2 Sensor		Result with UR-FALL (10 fold mean)			
	Yes or No Stream-1	No of GSTCN Skeleton	Yes or No Stream-2	Model Name	Accuracy	Precision	Recall	F1-score
1	No	-	Yes	Only CNN	97.78	93.79	92.92	93.02
2	No	-	Yes	Bi-LSTM with CNN	99.04	96.92	97.24	96.91
3	No	-	Yes	Bi-LSTM with Channel Attention	99.07	96.63	97.21	96.75
4	Yes	3	No	-	91.57	-	-	-
5	Yes	4	No	-	91.56	-	-	-
6	Yes	6	No	-	91.86	-	-	-
7	Yes	9	No	-	91.67	-	-	-
8	Yes	3	Yes	Bi-LSTM with Channel Attention	98.53	-	-	-
8	Yes	9	Yes	Bi-LSTM with Channel Attention	98.66	-	-	-
9	Yes	6	Yes	Bi-LSTM with Channel Attention	99.09	97.06	97.18	96.99

Table 3. Ablation study of the proposed model for UR Fall multi-modal dataset.

Ablation	Stream-1 No GSTCN	Stream-2 BiLSTM-CNN	Result with UR-FALL (10 fold mean)			
			Accuracy	Precision	Recall	F1-Score
1	3	1	99.14	99.06	99.041	99.04
2	4	1	99.15	99.19	98.81	98.99
3	5	1	99.24	99.12	99.19	99.15
4	6	1	99.16	99.20	98.48	98.82
5	9	1	**99.32**	99.23	99.19	99.21

3) Are there any limitations to using Alphapose? In occlusions, some skeleton keypoints may be missing. Does it happen? If yes please discuss

Author response: Thank you for raising your concern. We appreciate your observation regarding the potential limitations of using Alphapose, particularly in cases of occlusions where some skeleton key points may be missing. Indeed, occlusions can lead to missing or inaccurate key points in any pose estimation method, including Alphapose. However, we chose Alphapose primarily for its open-source nature, which offers more flexibility and ease of deployment compared to other solutions. Additionally, Alphapose excels in tracking multiple people within a given frame, making it a suitable choice for our multimodal fall detection framework.

While it's true that missing key points due to occlusions can occur, Alphapose's performance, in our experience, is quite robust in handling such cases. In situations where key points are occluded, we incorporate techniques like interpolation or rely on neighboring frame data to mitigate the impact of missing information. However, we acknowledge that this may not always be perfect, especially in complex scenarios with severe occlusion.

In our previous work, we observed similar results when comparing Alphapose to other methods like Mediapipe and OpenPose. Alphapose, however, was notably effective in reducing frame-skipping issues, which is crucial for maintaining continuity in pose estimation. While we plan to explore Mediapipe in future experiments for pose extraction, we have found Alphapose to be a highly suitable solution for this particular task.

---

4) Please explain Class wise precision, recall, and F1-score in the paper.

   Author response: Thank you very much for raising your concern. We apologize for the confusion. In our paper, we report fold-wise accuracy, precision, recall, and F1-score for the 10-fold cross-validation, rather than class-wise metrics. These metrics were calculated for each fold to evaluate the model's performance on different subsets of the data, providing a comprehensive assessment across multiple validation sets. We clarify these confusion in the revised manuscript. Please see the revised manuscript for details.

5) Are the results in Table 8. are taken from those papers or the authors have trained the models by themselves? Please explain which modalities are they using and also highligh you modality. This will make the results more comparable.

   Author response: Thank you very much for raising your concern.Thank you for your comment. In Table 8 (change in the revised munsricpts Table 5), we provide a state-of-the-art comparison, where the results presented are taken directly from the corresponding papers. We have not trained these models ourselves but have reported the performance metrics as provided in the referenced works. To clarify the modalities used, the studies we are comparing employ a variety of sensor modalities, including multi-sensor approaches (e.g., IMU, EEG, inertial sensors) and skeleton-based data. In our proposed model, we use a multimodal approach that integrates both sensor data and skeleton data, which allows for a more comprehensive representation of human motion. We ensured that this distinction is clearly highlighted in the revised manuscript to make the comparison more transparent and understandable for the readers.

6) Please compare results with the following paper as it is also using multimodal data.

Ha, T.V., Nguyen, H., Huynh, S.T., Nguyen, T.T., Nguyen, B.T. (2022). Fall Detection Using Multimodal Data. In: Þór Jónsson, B., *et al.* MultiMedia Modeling. MMM 2022. Lecture Notes in Computer Science, vol 13141. Springer, Cham. https://doi.org/10.1007/978-3-030-98358-1_31

Author response: Thank you very much for raising your concern. According to your concern, we updated our manuscript with your reference. Please see the revised manuscript for details.

10. Abdullah, F.; Jalal, A. Semantic segmentation based crowd tracking and anomaly detection via neuro-fuzzy classifier in smart surveillance system. Arabian Journal for Science and Engineering **2023**, 48, 2173–2190.
11. Ha, T.V.; Nguyen, H.; Huynh, S.T.; Nguyen, T.T.; Nguyen, B.T. Fall detection using multimodal data. In Proceedings of the International Conference on Multimedia Modeling. Springer, 2022, pp. 392–403.
12. Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.L.; Grundmann, M. Mediapipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214 **2020**.

7) Figure 1 is not discussed in the text. In Figure 4, the rotated BiLSTM may make more sense; otherwise, it would be a bit confusing because of the arrows. The skeleton in equation 11 should be defined before being used here. Reference is missing at line 52, and the dataset at lines 96, 120, 136,156, 158, please fix them. At line 371, the sensor should be fixed.

Author response: Thank you very much for raising your concern. According to your concern, we updated the manuscript. Please see the manuscript for details. Please see the Figure 4 and other part for details.

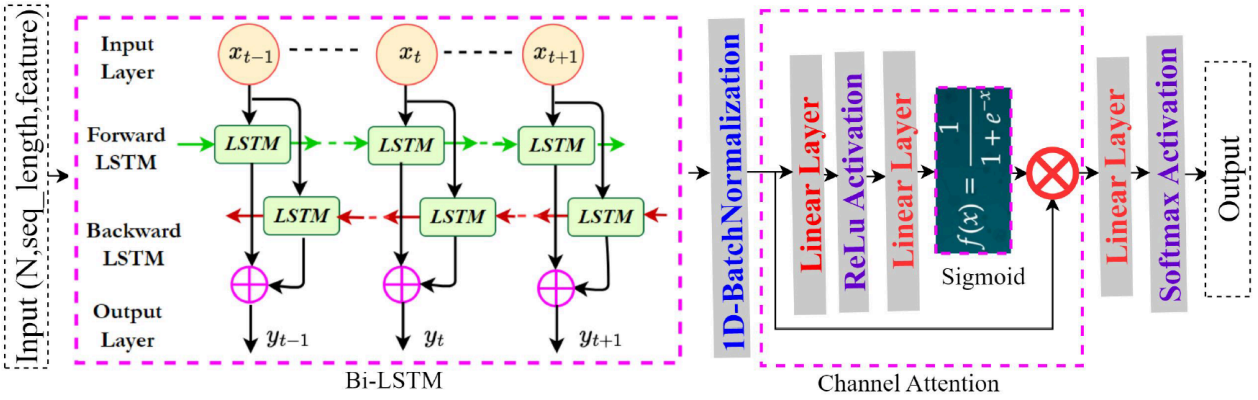


Figure 4. Integration of the CNN with Bi-LSTM model for sensor data modality feature extraction

8) The paper should be proofread for grammatical mistakes. Please fix the text on lines 174-176. At line 102, data sets -> datasets.

Author response: Thank you very much for raising your concern. According to your concern we updated the manuscript. Please see the manuscript for details.

Reviewer #2

Thank you for sharing this innovative multimodal fall detection framework. Seeing how you have integrated skeleton and sensor data to enhance accuracy and robustness is impressive. Using the Graph-based Spatial-Temporal Convolutional and Attention Neural Network (GSTCAN) alongside the Bi-LSTM with Channel Attention is exciting, as it effectively addresses existing systems' limitations.

Achieving a classification accuracy of 99.09% on the Fall Up dataset is an outstanding result, and it certainly highlights the potential of your system for improving fall detection and promoting safety among older people. It would be great to learn more about the practical applications of your framework and any plans for future developments or testing in real-world environments. Overall, this is a significant contribution to healthcare technology.

Here are some suggestions to improve the article:

1. Please arrange the keywords in alphabetical order.

   Author response: Thank you very much for raising your concern. Based on your concern, we have updated our revised manuscript. Please see the keywords section for details.

2. Please put a space before the in-text citations and correctly format them (Example: AlphaPose[51]---line no. 246) in the entire article.

   Author response: Thank you very much for raising your concern. According to your coner we updated our manuscript by including the space before the in-text citation in the entire article. Please see the revised manuscript for details.

3. The model is tested only on the Fall UP dataset, raising concerns about its generalizability across diverse datasets. Please evaluate datasets like UR Fall and SisFal to compare performance across different environments.

   Author response: Thank you very much for raising your concern.

Thank you for your valuable feedback. We appreciate your suggestion to evaluate our model on additional datasets such as UR Fall and SisFal to assess its generalizability across different environments. We did reach out to the dataset owners for permission to use the SisFal datasets, but unfortunately, we were unable to obtain the download permission in time for this study.

Based on your concern, we will evaluate the UR-Fall dataset. We include Table 3 for the ablation study with the UR-Fall dataset, Table 6 experimental results and Table 7 state-of-the-art comparison. Please see Section 5.5 and 5.6 for details about the UR-Fall dataset result.

Table 3. Ablation study of the proposed model for UR Fall multi-modal dataset.

Ablation	Stream-1 No GSTCN	Stream-2 BiLSTM-CNN	Result with UR-FALL (10 fold mean) Accuracy	Precision	Recall	F1-Score
1	3	1	99.14	99.06	99.041	99.04
2	4	1	99.15	99.19	98.81	98.99
3	5	1	99.24	99.12	99.19	99.15
4	6	1	99.16	99.20	98.48	98.82
5	9	1	**99.32**	99.23	99.19	99.21

Table 6. Cross Validation Fold wise precision, recall, and F1-score for UR-Fall multimodal dataset

Fold	Accuracy [%]	Precision [%]	Recall [%]	F1-Score [%]
1	100	100	100	100
2	99.31	99.53	98.70	99.11
3	100	100	100	100
4	99.68	99.80	99.23	99.51
5	96.69	96.68	96.71	96.68
6	99.68	99.17	99.80	99.48
7	100	100	100	100
8	99.42	98.65	99.63	99.13
9	100	100	100	100
10	98.38	98.47	97.83	98.14
Average	99.32	99.23	99.19	99.21

Table 7. State-of-the-Art Comparison of the Proposed Model on the UR Fall Multimodal Dataset

Author	Data Modality	Method Name	Accuracy [%]	Precision [%]	Recall [%]
Kwolek [46]	Depth	SVM	94.28	-	-
Youssfi [61]	Skeleton	SVM	96.55	-	-
Cai [62]	-	HCAE	90.50	-	-
Chen et al. [63]	RGB	Bi-LsTM	96.70	-	-
Zheng[51]	Skeleton		97.28	97.15	97.43
Wang[64]	Keypoints	-	97.33	97.78	97.78
Our Proposed System	Sensor+Skeleton	Two-Stream DNN	99.32	99.23	99.19

4. The paper acknowledges that fall incidents account for only 20% of the dataset, which could lead to biased results. You can use SMOTE (Synthetic Minority Over-sampling Technique) to generate additional fall samples.

   Author response: Thank you very much for raising your concern. We acknowledge the concern regarding the class imbalance in the dataset, where fall incidents constitute only 20% of the samples, which could lead to potential bias in the results. While we initially refrained from applying SMOTE due to the complexity of working with video data and maintaining temporal dependencies. In light of your suggestion, we plan to explore the application of SMOTE in future work for the video dataset to generate additional fall samples. However, we need to ensure that this technique is carefully adapted to preserve the temporal relationships inherent in video data and that the generated synthetic samples remain realistic and consistent with the task.

5. The proposed multimodal system, while effective, could be computationally expensive for real-time applications. Dimensionality Reduction techniques such as Principal Component Analysis (PCA) could be applied to remove redundant features from the skeleton and sensor

data. Also, Autoencoders could be used to compress sensor data before feeding it into Bi-LSTM.

Please look at the following article:

Ray, S.; Alshouiliy, K.; Agrawal, D.P. Dimensionality Reduction for Human Activity Recognition Using Google Colab. *Information* **2021**, 12, 6. https://doi.org/10.3390/info12010006

Author response: Thank you very much for raising your concern. We agree that dimensionality reduction techniques such as Principal Component Analysis (PCA) can be effective in reducing the computational cost by removing redundant features from the data. However, in our current setup, the sensor data consists of multiple channels, and selecting the most effective channels using a simple PCA approach may not always lead to optimal results. This is due to the complexity and variability in the sensor data across different contexts. We plan to explore more advanced channel selection techniques in the future, including PCA and other methods, to optimize the feature selection process. We appreciate your feedback and will incorporate these considerations into our future work.

45. Ray, S.; Alshouiliy, K.; Agrawal, D.P. Dimensionality reduction for human activity recognition using google colab. Information **2020**, 12, 6.
46. Kwolek, B.; Kepski, M. Human fall detection on embedded platform using depth maps and wireless accelerometer. Computer methods and programs in biomedicine **2014**, 117, 489–501.
47. Igual, R.; Medrano, C.; Plaza, I. Challenges, issues and trends in fall detection systems. Biomedical engineering online **2013**, 12, 66.
48. Zhang, Z.; Conly, C.; Athitsos, V. A survey on vision-based fall detection. In Proceedings of the Proceedings of the 8th ACM international conference on PErvasive technologies related to assistive environments, 2015, pp. 1–7.